《数据挖掘》课程期中测验

学号_____　　姓名_____　　得分_____

一、（20分）

某银行拟基于 200 万笔历史交易构建信用卡欺诈检测模型，已知欺诈交易占比约为 0.3%。请回答：

（1）说明为什么不能以准确率（Accuracy）作为主要评价指标，并用具体数值举例说明。（5分）

（2）设计一套完整的建模方案，须分别从数据层面、算法层面、评价层面各提出至少两项针对性措施，并简要说明其作用。（10分）

（3）模型上线后因误报过多导致客户投诉激增。此时应优先调整哪一环节？请结合精确率与召回率的权衡关系说明理由。（5分）

二、（25分）

某地铁车站附属结构的关键路径由三项连续工序组成：A 围护结构、B 基坑开挖、C 主体结构。项目经理按三点估算给出了每项工序的乐观值 a、最可能值 m 与悲观值 b（单位：天），见表 1。业主要求以 P90 工期作为节点考核依据，而项目部内部按均值（P50）承诺工期。

表 **1**　工序的三点估算（单位：天）

工序	乐观 a	最可能 m	悲观 b
A　围护结构	8	10	18
B　基坑开挖	11	14	23
C　主体结构	6	9	16

（1）三点估算（PERT，通过考虑最乐观、最可能和最悲观三种情景来得出估算结果，其核心是计算期望值）本身已有均值与方差的解析公式。请说明在什么条件下可以直接用解析法，本场景为何仍需蒙特卡洛，并指出解析法最容易失效的两个环节。（5分）

（2）三项工序由同一家分包商施工、共用同一批机械与班组。若按相互独立建模，会系统性低估风险。请说明这一偏差的方向与成因，并给出两种可操作的建模方案。（6分）

（3）设各工序服从三角分布，请按均值 =（a + m + b）/ 3、方差 =（$a^2 + m^2 + b^2 - am - ab - mb$）/ 18，计算三项工序的均值与标准差，求总工期的均值与标准差（独立假设），并用正态近似求 P90 工期（结果保留三位小数）。（6分）

（4）某次 10000 次模拟显示总工期均值的标准误为 0.04 天。若要把标准误降到 0.01 天，需要增加多少次模拟？除单纯增加次数外，请再给出两种降低方差的抽样技术，并各用一句话说明其原理。（5分）

（5）项目部提出"用均值 38.3 天加 5 天固定安全量作为考核工期"。请评价该做法；若蒙特卡洛给出的 P90 为 43.6 天、P99 为 48.0 天，而正态近似分别为 43.39 天与 47.51 天，请说明两组数字的差异说明了什么。（3分）

三、（25分）

某内容平台约有 240 万注册用户、8600 万条有向关注边，平均每个用户关注 36 人、约有 12 个粉丝。平台计划在新品发布期挑选 500 名"种子用户"寄送样品并激励其发布内容，期望通过其社交关系带动二次传播。

（1）度中心性、介数中心性、接近中心性、特征向量（或 PageRank，网页级别评估指标）中心性分别刻画了"重要性"的哪个侧面？请说明为什么"粉丝数最多"的用户未必是合适的种子，至少给出三条理由。（6分）

（2）表 2 是一个含 5 个节点的有向关注网络（箭头由关注者指向被关注者）。设阻尼系数 $d = 0.85$，初始 PageRank 值均为 0.2。写出 PageRank 的迭代公式，计算前两轮迭代结果（保留四位小数），并说明最终排序是否支持"入度越高者排名必高"。（7分）

表 **2**　有向关注关系

用户	其关注的用户	用户	其关注的用户
A	B、C	D	A
B	C	E	D
C	A	—	—

（3）若用传染病模型刻画内容扩散，SIR 与独立级联（IC）模型各自的假设是什么？请说明参数（β、γ 或边级传播概率 p）在真实数据中如何估计，以及为什么不能只报告一个 R_0。（5分）

（4）用 Louvain 等算法（即社区检测的经典算法，通过最大化模块度来识别网络中紧密连接的子群体），做社区划分后，有观点主张"500 个种子应全部投向最大的社区"。请评价这一主张，并说明跨社区分散播种的收益与代价。（4分）

（5）投放结束后，数据显示种子用户社交圈的购买率为 9.8%，平台平均为 4.1%，团队据此认定"种子策略带来 5.7 个百分点提升"。请指出该推断的问题，并给出可执行的改进设计。（3分）

四、（30分）

某量化团队把每只股票的 20 个交易日 K 线，渲染为 32×32 的灰度图像（横轴为时间、纵轴为价格、亮度为成交量），用二维卷积网络预测未来 5 个交易日的涨跌方向，网络结构见表 3。团队声称"卷积网络的平移不变性，天然适合识别 K 线形态"。

表 **3**　网络结构配置

层	配置
输入	32×32×1 灰度图
Conv1	3×3 核，32 个通道，步长 1，填充 1
MaxPool1	2×2，步长 2
Conv2	3×3 核，64 个通道，步长 1，填充 1
MaxPool2	2×2，步长 2
Flatten 后接 FC	展平后接 128 个神经元的全连接层，再输出 1 个标量

（1）取图像左上角的 3×3 像素块（见表 4）作为输入，用 2×2 的卷积核（见表 4）做一次 valid 卷积（即步长 1、无填充、偏置为 0），写出 2×2 的输出矩阵；再经 ReLU 激活（即正数原样输出，负数归零），最后做 2×2 的最大池化与平均池化，分别给出两个标量结果。（7分）

表 **4**　输入块与卷积核

输入块 X（3×3）	卷积核 K（2×2）
1　2　0 0　1　3 2　1　1	1　0 −1　1

（2）逐层写出各卷积层与池化层的输出张量尺寸，并计算卷积层与全连接层各自的参数量。据此指出该结构的参数分布存在什么问题，给出一种能把总参数量降低一个数量级的改法，并说明改造后的参数总量。（7分）

（3）该团队认为"卷积的平移不变性天然适合识别 K 线形态"。请分别就图像的横向（时间）与纵向（价格）两个方向，判断平移不变性假设是否成立，并说明若假设不成立会导致什么后果、可用哪些手段补救。（6分）

（4）从数值表示的角度，分析"把 K 线渲染为 32×32 灰度图"相对"直接在原始价格序列上建模"损失了哪些信息。并说明为什么用 OHLC 四通道比用 RGB 三通道更合理。其中：OHLC 是金融数据的一种结构化表示（开盘、最高、最低、收盘四个价格），而 RGB 是颜色编码方式（红、绿、蓝三个通道）。（6分）

（5）该团队计划加载 ImageNet 预训练权重（即在大规模 ImageNet 数据集上训练好的模型参数，可直接用于迁移学习，省去从头训练的时间和数据成本）做迁移，并沿用图像领域的翻转、旋转类数据增强。请分别判断这两种做法在该任务上是否可行，并各给出一条理由。（4分）